The accuracy of Apple Watch measurements: a living systematic review and meta-analysis.

Supplementary Information (S.I.)

Table of contents

Supplementary Note 1: Additional results synthesis

Risk of bias

Six heart rate studies were rated as 'low' risk, whereas 11 were rated as 'some concerns' and 21 as 'high'. Risk of bias most frequently arose from inappropriate statistical analysis (Domain 4), such as not accounting for repeated measures or limited description of analysis methodology. Ten (26%) studies did not appropriately select participants to represent the desired target population (Domain 1). Eighteen studies (47%) did not state that manufacturer guidelines for device wear were followed, as recommended by INTERLIVE, and were rated as 'some concerns' for Domain 2. In contrast, 35 (92%) studies were rated as low risk for Domain 3 (reference standard).

To assess the robustness of our meta-analysis, we conducted sensitivity analysis of heart rate measurements with rest and exercise conditions combined, excluding studies rated as 'high' risk of bias. Mean bias (MB) and limits of agreement (LoA) were comparable to our primary analysis (MB -0.50 bpm [95% CI -1.47–0.47]; LoA -7.54 to 6.53; 13 studies; figure S1).

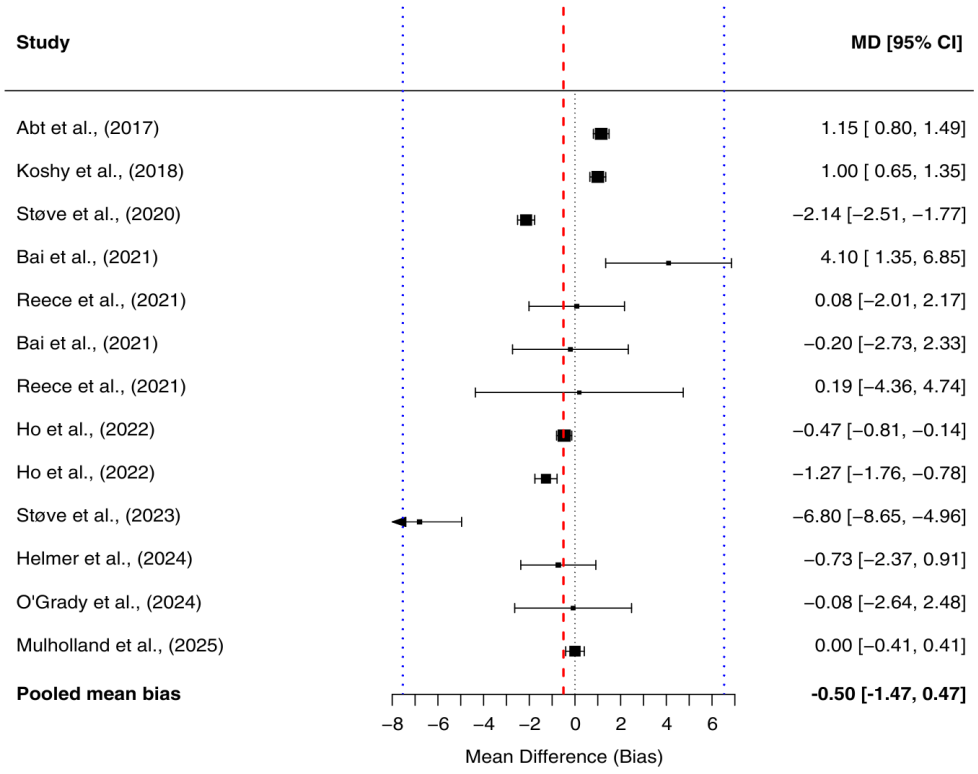

Figure S1. Forest plot for sensitivity analysis of heart rate under all conditions, excluding studies rated at high risk of bias. The dashed red line represents the pooled mean bias, and the dashed blue line represents the pooled limits of agreement.

No studies which validated atrial fibrillation detection were rated 'low' risk (four [24%] some concerns; 13 [76%] high). Many of these studies included disproportionately few

females and excluded uninterpretable ECG recordings from all analyses. Reference to manufacturer guidelines for device wear, or specific detail regarding device placement on the wrist, was often lacking (Domain 2). This was also the case for studies validating ECG waveform morphology, where two (29%) and five (71%) studies were rated as 'some concerns' and 'high', respectively.

For energy expenditure, three of the seven studies did not validate Apple Watch in a balanced sample, representative of the defined target population. Similarly, over 70% of participants were male in four (40%) blood oxygen saturation studies. Four (40%) blood oxygen saturation studies did not appropriately report missing data, and two (20%) excluded all unsuccessful measurements from analysis. Of studies evaluating heart rate variability, hypertension notification and sleep apnoea, none were rated as high risk. The ratings for each study across all subdomains are available as a supplementary file.

Heart rate

The lowest agreement for heart rate during exercise was reported in a study involving barbell resistance exercises, where movement patterns were irregular, and heart rate changed quickly[1]. Three studies reported decreased accuracy with increased arm motion[1-3]. Few studies reported sampling frequency relative to the criterion. Among cardiac populations, agreement was lower for patients in atrial fibrillation than those in sinus rhythm, with mean difference of up to 8 bpm reported. There was no evident trend in accuracy between different Apple Watch models, and just five studies validated Series 5 or later.

To compare findings across different Apple Watch models, we performed exploratory subgroup analysis according to the generation of optical heart rate sensor: first-generation (Apple Watch models up to Series 3), second-generation (Series 4–5 and all SE models), and third-generation (Series 6 onwards, including Ultra models). We included heart rate measured under all conditions, as in the primary analysis. Our results showed narrower limits of agreement for the third-generation sensor compared to our primary analysis (LoA -3.68 to 2.59; 8 studies; figure S2). Contrastingly, the limits of agreement were wider for the first- and second-generation sensors when compared to our primary analysis (first-generation: LoA -10.62 to 11.26, 9 studies; second generation: LoA -11.39 to 12.30, 5 studies). Across all sensor generations, mean bias was similar to that found in our primary meta-analysis. Results are presented in Table S1.

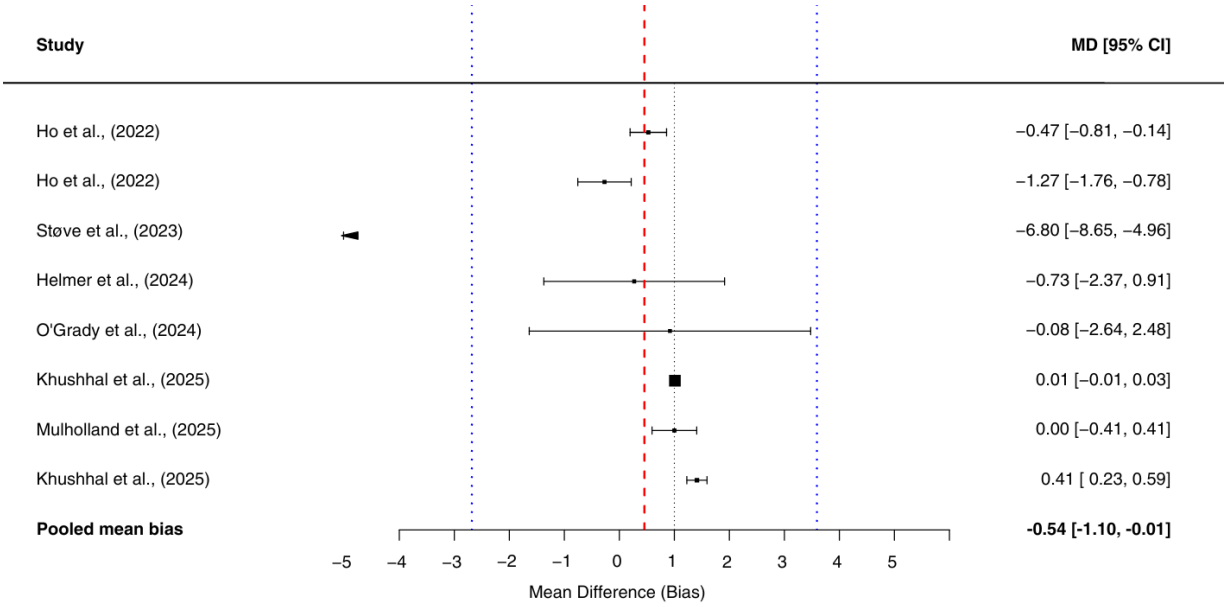

Figure S2. Forest plot for heart rate under all conditions, with 3rd generation optical heart rate sensor. The red line depicts pooled mean bias, while the blue lines show pooled Bland-Altman limits of agreement.

The final subgroup analysis for heart rate included results from patient cohorts only, across resting and exercise conditions (figure S3). Seven studies were eligible, and the pooled mean bias was 0.36 bpm (LoA -5.08 to 5.80 [95% CI -13.21–13.93]).

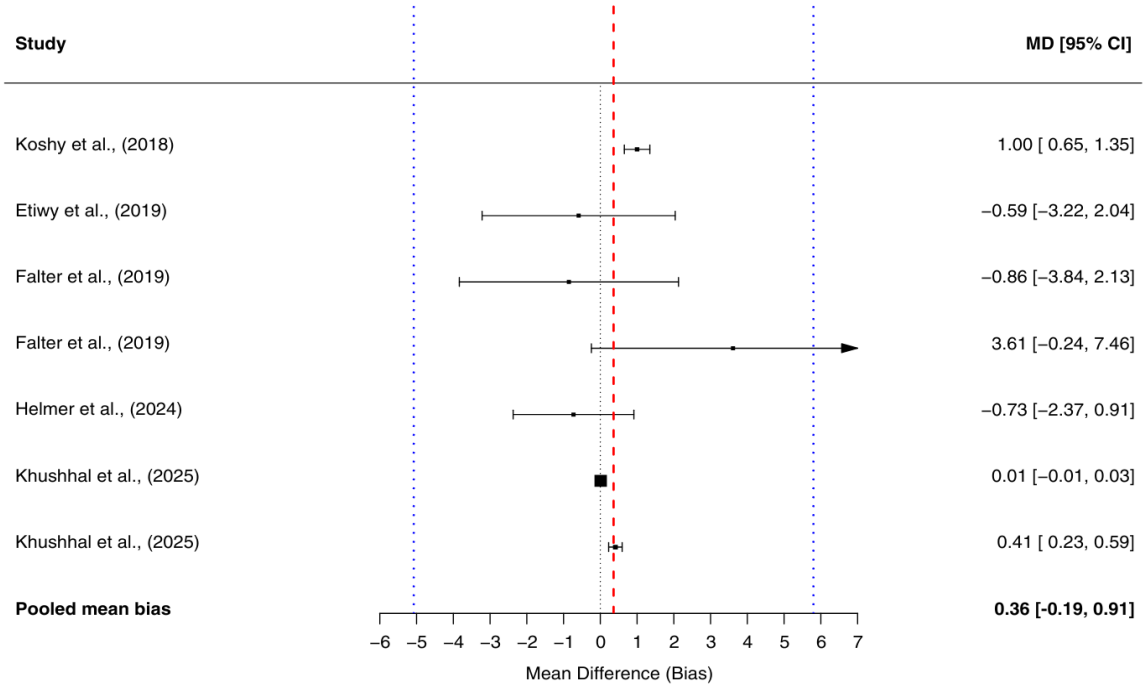

Figure S3. Forest plot for heart rate under all conditions among patient cohorts only. The red line depicts pooled mean bias, while the blue lines show pooled Bland-Altman limits of agreement.

To assess publication bias, we generated a funnel plot (figure S4) to visually inspect for potential small-study effects. Formal statistical testing was not conducted because standard methods, such as Egger's test, are designed for effect measures and rely on assumptions (e.g., a linear relationship between effect size and standard error) that do not hold for mean difference and limits of agreement data. Upon visual inspection, there was no clear evidence of asymmetry, despite a small number of outliers, suggesting a low likelihood of publication bias.

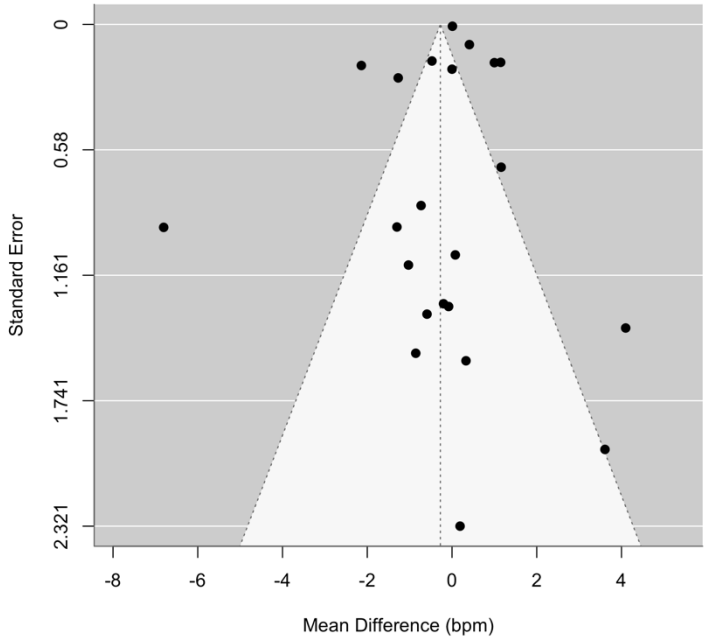

Figure S4. Funnel plot of heart rate showing standard error against mean difference, in beats per minute.

Heart Rate Variability

Heart rate variability was validated by one study (n=39)[4]. Mean absolute error was 28.9%, and measurements exceeded the equivalence threshold.

Atrial fibrillation detection

Pooled false positive rate was 0.094 (95% CI 0.04–0.19). The rate of inconclusive tracings was between 15 and 25% in several studies[5-10]. Almost a third of studies excluded inconclusive tracings from analyses entirely, yielding results that may not represent real-world use. In addition, diagnostic accuracy contingency tables were reported very infrequently. This veiled the calculation of sensitivity and specificity values. Consequently, no study was rated as low risk of bias.

One study enrolled over 400,000 participants, but fewer than 0.5% contributed to analyses[11]. In that study, 86 participants received an irregular pulse notification while wearing an ECG patch (criterion measure), yielding a positive predictive value of 84%.

However, atrial fibrillation was confirmed in only 34% of the 450 participants who received an irregular pulse notification at any time – whilst simultaneously wearing an ECG patch or not.

We conducted subgroup analysis to examine the influence of hardware and software version on atrial fibrillation detection performance. One group included Apple Watch Series 4 and 5, whereas the other included Series 6 (all studies in our review validated either Series 4, 5, or 6). This grouping was chosen as the third-generation optical heart rate sensor was introduced with Series 6, and the ECG app was updated from version 1.0 to 2.0 two months after its release, via watchOS 7.2. Compared to the older models, we found higher sensitivity for the Series 6 subgroup (new: 0.90 [95% CI 0.65–0.98], 5 studies; old: 0.73 [95% CI 0.41–0.91, 6 studies]). In contrast, specificity was lower for the Series 6 subgroup (new: 0.87 [95%CI 0.68–0.95]; old: 0.95 [95% CI 0.80–0.99]).

Hypertension notification

Apple conducted a clinical validation study prior to the release of their hypertension notification (HTNF) feature in September 2025, which was submitted for FDA 510(k) approval and reported in a white paper. 2,229 (1,268 male [57%]) participants were recruited, of which 1,863 were included in the final analysis. A diverse cohort of various ethnicities and BMI ranges, with and without hypertension were recruited. Sensitivity and specificity were determined based on average blood pressure readings taken twice daily at home over 30 days using the criterion, OMRON Evolv Wireless Upper Arm Blood Pressure Monitor. Participants were instructed to wear an Apple Watch each day to collect PPG data which was processed by the hypertension notification algorithm. Overall sensitivity was 41.2% (95% CI [37.2%, 45.3%]), and specificity was 92.3% (95% CI [90.6%, 93.7%]). Specificity among those with normal blood pressure was 95.3% (95% CI [93.7%, 96.5%]), and the sensitivity for Stage 2 hypertension, which is more severe and requires more aggressive management, was 53.7% (95% CI [47.7%, 59.7%]). No independent validation studies have been conducted to date.

Blood oxygen saturation

Rajakariar and colleagues suggested that Apple Watch did not demonstrate sufficient accuracy for reliable detection of hypoxemia. Several studies found rates of unsuccessful measurements of approximately 5%[12-16], and only one study included missing data in analysis[15]. Upon visual inspection of funnel plots, there was no clear evidence of asymmetry, indicating low likelihood of publication bias (figure S5).

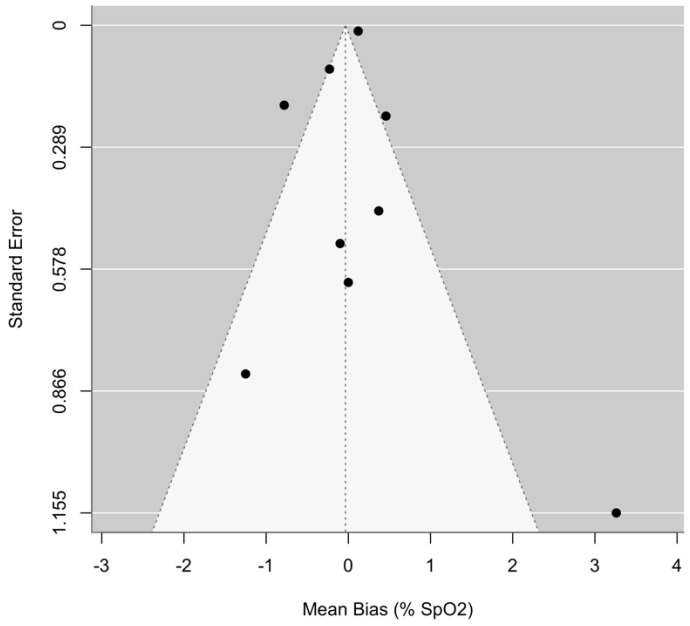

Figure S5. Funnel plot of blood oxygen saturation showing standard error against mean bias, expressed as percentage SpO_2.

Due to differences in measurement accuracy noted during narrative results synthesis, we conducted subgroup analysis comparing blood oxygen saturation measurements in normoxic ranges to those in hypoxic ranges (<94% SpO_2)[17]. In hypoxic ranges, limits of agreement were notably wider in comparison to our primary analysis (hypoxic MB 0.43% [95% CI -3.85–4.71]; LoA -8.35 to 9.21), and the mean bias was positive rather than negative, indicating overestimation relative to the criterion.

Sound exposure

To validate sound exposure, Fischer and colleagues compared Apple Watch Series 6 to a Class 1 sound level metre in various public environments[18]. A minor underestimation (0.5 dBA) by Apple Watch was found overall.

Six-minute walk test distance estimation

Apple's white paper describing development and validation of their six-minute walk test distance estimation feature allocated 449 adults (184 males [41%]) aged 65 or older to the validation group. The distance obtained during supervised six-minute walk tests was the criterion measure. Although limited detail of results is presented, at a threshold of 360 metres, the Area Under the Curve was 0.946, sensitivity was 80.8%, and specificity was 90.5%. Error, described as weekly e6MWD minus the nearest reference test, was 1 metre (SD 55). There were no independent validation studies eligible for inclusion in this review.

Table S1. Results of meta-analyses, sensitivity analysis, and subgroup analyses for heart rate and blood oxygen saturation.

Number of studies	Pooled mean bias (95% CI)	Standard deviation	Tau^2	Limits of agreement (lower, upper)	CI for limits of agreement (robust variance estimate; moderator model)
Heart rate overall					
22	-0.274 (-0.72–0.17)	3.80	0.529	-7.19 to 6.64	-10.62, 10.07; -11.96, 11.42
Heart rate overall, excluding studies at high risk of bias					
13	-0.503 (-1.47–0.47)	3.23	1.94	-7.54 to 6.53	-11.30, 10.29; -10.92, 9.92
Resting heart rate					
9	0.213 (-0.65–1.07)	4.09	0.67	-8.14 to 8.56	-14.19, 14.61; -14.66, 15.09
Heart rate during exercise					
13	-0.627 (-1.37–0.12)	2.96	0.93	-6.86 to 5.60	-11.36, 10.10; -13.36, 12.11
Heart rate, including only patient populations					
7	0.361 (-3.85–4.71)	2.68	0.16	-5.08 to 5.80	-13.21, 13.93; -15.12, 15.84
Heart rate measured with the 3rd generation optical heart rate sensor					
8	-0.545 (-1.08 to -0.01)	1.48	0.26	-3.68 to 2.59	-6.78, 5.69; -7.60, 6.51
Heart rate measured with the 2nd generation optical heart rate sensor					
5	0.452 (-0.85–1.75)	5.92	0.00	-11.39 to 12.30	-17.98, 18.88; -17.95, 18.85
Heart rate measured with the 1st generation optical heart rate sensor					
9	0.319 (-1.26–1.90)	5.18	3.14	-10.62 to 11.26	-18.22, 18.86; -18.21, 18.85
Blood oxygen saturation					
9	-0.035 (-0.42–0.35)	1.95	0.13	-4.01 to 3.94	-6.56, 6.49; -7.54, 7.47
Blood oxygen saturation, including only measurements in hypoxic ranges					
5	0.430 (-3.85–4.71)	2.79	11.46	-8.35 to 9.21	-63.02, 63.88; -63.04, 63.90

Meta-analysis of atrial fibrillation detection
Bivariate diagnostic random-effects meta-analysis.
Estimation method: REML

Fixed effects coefficients

	Estimate	Std. Error	z	Pr(>|z|)	95% CI
tsens. (Intercept)	1.336	0.462	2.888	0.004	0.429 to 2.242
tfpr. (Intercept)	-2.270	0.416	-5.450	0.000	-3.086 to -1.454
Sensitivity	0.792	–	–	–	0.606 to 0.904
False positive rate	0.094	–	–	–	0.044 to 0.189

Variance components: between-studies Std. Dev and correlation matrix

	Std. Dev	tsens	tfpr
tsens	1.433	1.000	–
tfpr	1.267	0.050	1.000

logLik	AIC	BIC
17.718	-25.435	-19.980

Area Under the Curve (AUC): 0.925
Partial AUC (restricted to observed FPRs and normalized): 0.793

I^2 estimates:
Zhou and Dendukuri approach: 54.9%
Holling sample size unadjusted approaches: 62.3–91.8%
Holling sample size adjusted approaches: 3.5–7.7%

Supplementary Note 2: Search strategy

The following search strategy was used in each database:

("Apple Watch" OR "Applewatch" OR "iwatch")
AND
(valid* OR accur* OR reliab* OR compar* OR consist* OR sensitiv* OR specific* OR reproduc*)

PubMed, Embase, Web of Science, Google Scholar: Search included 'all fields'
SPORTDiscus: Tx All Text
IEEE Xplore: All
CINAHL: All text
Cochrane Library, Scopus: Title, abstract, keywords

The US Food and Drug Administration 510(k) Premarket Notification Database was searched (https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm) by entering 'Apple Inc.' in the Applicant Name field. All other fields were left blank. For each result, the Summary letter was read in full.

The Apple Health website (apple.com/health) was manually searched.

Supplementary Note 3: Description of criterion measures

Assessment of Apple Watch measurement accuracy hinges on the validity of the criterion measure used. We adopted a stringent approach, where gold standard methods of assessment or validated criterion measures were deemed acceptable. During our screening process, the criterion measure used in each study was evaluated independently to determine its suitability. Below, we describe the criterion measures we accepted for each health-related metric.

Heart rate: 12 or 6-lead electrocardiography, Biopac ECG or similar, chest strap using electrodes for measurement (e.g. Polar H10).

Heart rate variability: We accepted the same hardware as for heart rate, with the addition of a validated HRV software, such as Kubios HRV.

Atrial fibrillation detection: 12 or 6-lead electrocardiography and Biopac ECG or similar, with the addition of clinician-interpreted diagnosis. Validated insertable cardiac monitors or ECG patches.

ECG waveform morphology: 12 or 6-lead electrocardiography, Biopac ECG or similar, with the addition of clinician-based measurement of intervals, or measurement with a validated software.

Step count and wheelchair push count: Direct observation and manual counting. Research-grade wearables were not accepted for step count due to conflicting evidence regarding their validity. The authorship team thoroughly evaluated the validation literature of research-grade wearables, finding that some studies reported acceptable validity, whereas others did not. One study, which compared Apple Watch and wrist, hip and ankle, research-grade devices to direct observation, found lower error for Apple Watch than for the research-grade device[19]. Due to this contrasting evidence, they were excluded as criterion methods.

Energy expenditure: Indirect calorimetry (e.g. Cosmed Quark CPET), validated portable indirect calorimetry (e.g. Cosmed K5), doubly-labelled water. Research-grade wrist-worn wearables, such as Actigraph, were not accepted.

SpO_2: Arterial blood gas, or medical-grade pulse oximetry.

VO_2 max: Indirect calorimetry (e.g. Cosmed Quark CPET), or validated portable indirect calorimetry (e.g. Cosmed K5).

Sleep: Polysomnography.

Sleep apnea: Polysomnography or a clinically validated home sleep apnea test kit.

Sound exposure: Class 1 sound level meter.

Hypertension notification: Blood pressure monitor with FDA approval as a Class II medical device, MDR CE-marking, or ISO/AAMI/ESH standards.

Supplementary Note 4: Wearing Apple Watch appropriately

Guidelines for appropriately wearing Apple Watch are provided by Apple (https://support.apple.com/en-ie/118234), and each study included in our review was required to adhere to them. Users are instructed to wear Apple Watch snugly, but comfortably, just proximal to the ulnar and radial styloid processes. Wearing Apple Watch loosely, on or distal to the styloid processes, or at anatomical locations other than the wrist is regarded as inappropriate.

We did not include studies in which more than one PPG-based wearable was worn on the same wrist. The majority of Apple Watch health metrics are derived, at least in part, from photoplethysmography (PPG), in which green or red LED light is projected into the skin to detect volumetric changes in blood flow. PPG signals are highly sensitive to light interference and motion artefact. When two wearables are placed on the same wrist, this can cause interference between light sources and reflected signals, thereby distorting the PPG waveform and compromising measurement accuracy. Even ambient light has been shown to degrade PPG signal quality. In addition, wearing multiple devices on the same wrist inevitably alters device placement, affects sensor–skin contact, and increases distal mass and motion artefact, all of which can impair the function of both devices.

Supplementary Note 5: Risk of bias tool – adapted COSMIN

A detailed risk of bias table, with individual domain results per study, is available in Supplementary Data 1.

Domain 1: Participants

1.1: Were participants selected appropriately to represent the desired target population defined by the study authors?
1.2: Did all, or nearly all, participants sampled for the study contribute with data to be included in the analysis of criterion validity?

Domain 2: Index measure (Apple Watch)

2.1: Was the index measure (Apple Watch) administered during the sampling of data according to the manufacturer's instructions?

Domain 3: Reference standard (criterion measure)

3.1: Was the criterion measure instrument to assess the health measurement appropriate and administered appropriately during the test?
3.2: Was the test protocol (including criteria for a valid test) for assessment appropriate?
3.3: Was the assessor blinded to the measurement of the index measure? Did the assessor administer the criterion measure test without knowledge of values from the index measure?
3.4: Was the time interval between the criterion and index measurements appropriate?

Domain 4: Statistical analysis

4.1: Was the statistical approach to estimate agreement appropriate?

Supplementary Note 6: Back-calculation of statistics

For our chosen methods of meta-analyses – the framework for Bland-Altman meta-analysis proposed by Tipton & Shuster (2017), and bivariate meta-analysis of sensitivity and specificity, proposed by Reitsma et al., (2005) – we required specific statistical information from each individual study in order to include it.

Back-calculating standard deviation of the differences

For Bland-Altman meta-analysis of heart rate and blood oxygen saturation, according to Tipton & Shuster, we required a mean difference value along with the standard deviation of the differences. Many studies, however, did not report the standard deviation of the differences. It is possible to back-calculate the standard deviation of the differences from the 95% limits of agreement by rearranging the formula used to compute the limits of agreement, given as:

$$LoA = mean\ difference \pm 1.96 \times s_d$$

where s_d is the standard deviation of the differences.

The rearranged formula is given as:

$$s_d = \frac{Upper\ LoA - Lower\ LoA}{3.92}$$

Back-calculating 2x2 diagnostic accuracy tables

For bivariate meta-analysis of sensitivity and specificity for atrial fibrillation detection, according to Reitsma et al. (2005), a 2x2 diagnostic accuracy table was required. This involved reporting the number of true positives and negatives, and false positives and negatives. Very few studies did so, however. Using the methods described by Taylor and colleagues, we back-calculated these statistics. To achieve this, each study had to report: 1) the number of participants with and without atrial fibrillation in their sample, 2) the total sample size, 3) the overall sensitivity and specificity values. Taylor and colleagues state that, "The four numbers (i.e., TP, FN, TN, FP) may be calculated from the test's sensitivity and specificity, the prevalence of the condition, and the total number of participants tested using equations 1–4". The equations are provided in Table 1 and Figure 2 in their manuscript, which is accessible at: https://doi.org/10.1136/bmjebm-2020-111650

Supplementary Note 7: Pooling mean difference values

To prevent unit-of-analysis errors, we included only one estimate per study per condition in meta-analyses of heart rate and blood oxygen saturation. A number of studies reported multiple estimates per condition – one each for exercise at low, moderate, or vigorous intensity, for instance. Borenstein and colleagues (2009) note that including multiple estimates from the same study as if they were independent violates the assumption of independence, potentially giving the study undue weight and underestimating the precision of the summary effect. To address this, we combined multiple estimates within each study using the fixed-effect inverse-variance method outlined by Borenstein and colleagues to generate a single pooled estimate per study.

The formula for weights (w_{ij}), for each condition i within study j was given as:

$$w_{ij} = \frac{1}{Var(d_{ij})}$$

The formula for pooled mean difference was:

$$\hat{d}j = \frac{\sum_i wijd_{ij}}{\sum_i w_{ij}}$$

We calculated within-study variance as:

$$Var(d_{ij}) = \frac{SD_{ij}^2}{n_{ij}}$$

Supplementary Note 8: PRISMA checklist

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title, page 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract, page 1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Page 3
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Page 4
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Page 13
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Page 12
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Page 12; Supplementary Information Note 2
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Page 12–13
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Page 13
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Page 13
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Page 13
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Page 13–14
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Page 13
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Page 14
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Page 14
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Tables and figures cited in Results text
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Page 14
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Page 14
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Described in Results (Page 5) –

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
			post hoc testing
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Page 13–14
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Page 13–14
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Page 4
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	No such studies included
Study characteristics	17	Cite each included study and present its characteristics.	Table 1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Page 4; Supplementary Risk of Bias File
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Fig 2, 3, 4, 5
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Page 4
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Page 5–6
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Page 5–6
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Page 5
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Page 4
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Page 5
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Page 9
	23b	Discuss any limitations of the evidence included in the review.	Page 10
	23c	Discuss any limitations of the review processes used.	Page 10–11
	23d	Discuss implications of the results for practice, policy, and future research.	Page 11–12
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Page 12
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Page 12
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	No amendments made
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Page 15
Competing interests	26	Declare any competing interests of review authors.	Page 15
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Page 15

Supplementary Note 9: Studies excluded during full-text review

First author (year)	Study title	Reason for exclusion
Abi-Mansour et al., (2024)	Accuracy and role of consumer facing wearable technology for continuous monitoring during endoscopic procedures	Inappropriate intervention. Smartwatch monitoring was 'augmented'.
Abou et al., (2024)	Sensitivity of Apple Watch fall detection feature among wheelchair users	Falls were premeditated in lab conditions rather than true falls.
Abt et al., (2018)	Measuring Moderate-Intensity Exercise with the Apple Watch: Validation Study	Inadequate description of intervention, criterion measure, and statistical analysis.
Achermann et al., (2023)	Velocity-Based Strength Training: The Validity and Personal Monitoring of Barbell Velocity with the Apple Watch	Inappropriate metric.
Aggarwal et al., (2020)	Measurement of Corrected QT Interval on Apple Watch Series 4 Electrocardiogram Has Moderate to Strong Agreement with 12-lead Electrocardiogram	Inadequate description of intervention and analysis.
Ahluwalia et al., (2024)	Validation of diagnostic wearable ECG data from high-risk patients and integration into the atrial fibrillation catheter ablation pathway	Inappropriate criterion measure (clinician assessment of Apple Watch ECG vs automated Apple Watch classification.
Al-Kaisey et al., (2019)	Accuracy of wrist-worn heart rate monitors for chronotropic assessment in atrial fibrillation	Manual duplicate removal.
Alnasser et al., (2023)	The Reliability of the Apple Watch's Electrocardiogram	Inadequate description of statistical methods.
Alzahrani et al., (2025)	The use of Apple smartwatches to obtain vital signs readings in surgical patients	Inadequate reporting. Insufficient description of criterion measure and testing protocol, and statistical approaches. Inappropriate pooling of data in analyses.
Apple (2021)	Using Apple Watch to Estimate Cardio Fitness with VO_2 max	Inadequate reporting. The criterion method is not adequately described,

		nor is the procedure for determining that true VO_2 max was attained.
Apple (2024)	Using Apple Watch to measure heart rate, calorimetry, and activity	While this white paper provides detail on feature development and method of measurement, it doesn't provide detail on validation.
Arslan et al., (2024)	Accuracy of the Apple Watch in measuring oxygen saturation: comparison with pulse oximetry and ABG	Manual duplicate removal.
Auepanwiriyakul et al., (2020)	Accuracy and Acceptability of Wearable Motion Tracking for Inpatient Monitoring Using Smartwatches	Inappropriate metric – motion analysis.
Avidan et al., (2025)	From Wrist to Precision: Enhanced Atrial Flutter Detection with Modified Smartwatch Single-Lead Electrocardiogram Placement	Incorrect study objective. Evaluates the ability of clinicians to classify ECG recordings, rather than automatic classification by Apple Watch.
Bai et al., (2017)	Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR	Fitbit and Apple Watch worn concurrently on the left wrist.
Benning et al., (2020)	Comparison of accuracy of activity measurements with wearable activity trackers in wheelchair users: a preliminary evaluation	Multiple devices worn concurrently on the same wrist (non-dominant arm).
Binsch et al., (2016)	Comparison of three different physiological wristband sensor systems and their applicability for resilience-and work load monitoring	Apple Watch not tested appropriately.
Bonnelval et al., (2025)	Validity of Heart Rate Variability Measured with Apple Watch Series 6 Compared to Laboratory Measures	Inappropriate metric. RR and NN intervals were validated, following export of raw data, rather than natively available HRV SDNN values.
Boudreaux et al., (2018)	Validity of Wearable Activity Monitors during Cycling and Resistance Exercise	Inappropriate intervention. Eight devices worn simultaneously.
Bradel et al., (2023)	A comparison of the Apple Wearable device to implantable cardiac monitors	Inadequate information for inclusion (grey literature).
Breteler et al., (2019)	Measuring free-living physical activity with three commercially	Inappropriate criterion measure.

	available activity monitors for telemonitoring purposes: Validation study	
Briosa E Gala et al., (2024)	Single-lead ECGs with wearable technology: Diagnostic accuracy in patients with cardiovascular disease	Manual duplicate removal.
Bumgarner et al., (2018)	Smartwatch Algorithm for Automated Detection of Atrial Fibrillation	Inappropriate metric. Not available natively on Apple Watch.
Burke et al., (2023)	NOVEL AI ALGORITHM IMPROVES THE AUTOMATED DETECTION OF ATRIAL ARRHYTHMIAS FROM THE APPLE WATCH	Inadequate description of criterion measure.
Butler et al., (2023)	FEASIBILITY OF REMOTE MONITORING FOR FATAL CORONARY HEART DISEASE FROM SINGLE LEAD ECG	Incorrect study objective.
Caserman et al., (2024)	Assessing the Accuracy of Smartwatch-Based Estimation of Maximum Oxygen Uptake Using the Apple Watch Series 7: Validation Study.	Inappropriate criterion measure and inappropriate comparison of different exercise modalities. There was a lack of rigorous validation studies to ensure the validity of the portable indirect calorimetry used in this study. Additionally, authors compared a cycling protocol (criterion) to running / hiking / walking-derived estimates from Apple Watch. Comparison between these distinct exercise modalities produces inherently different results, regardless of measurement instrument.
Chase et al., (2024)	Accuracy of the Apple Watch Series 9 for Measures of Energy Expenditure and Heart Rate at Rest and During Exercise: Impact of Skin Pigmentation.	Inappropriate statistical analysis, with inadequate reporting.
Child et al., (2023)	Device Comparison and Needs Assessment of Pulse Oximetry Monitoring During Exercise for Patients With Interstitial Lung Diseases	Inadequate reporting (grey literature).

Chowdhury et al., (2017)	Assessment of laboratory and daily energy expenditure estimates from consumer multisensor physical activity monitors	Two wrist-worn devices worn simultaneously on a single wrist.
Danielsson et al., (2024)	Accuracy of the Apple Watch Series 4 and Fitbit Versa for Assessing Energy Expenditure and Heart Rate of Wheelchair Users During Treadmill Wheelchair Propulsion: Cross-sectional Study.	Multiple wrist-worn devices worn on a single wrist.
Delinière et al., (2024)	Wearable electrocardiogram devices in patients with congenital long QT syndrome: The SMART-QT study.	Inappropriate intervention. ECG in those younger than 22 years old.
Doggart et al., (2024)	Enhanced diagnostic accuracy of personal ECG recordings for atrial fibrillation through repeat measurements: A comparative analysis of Apple ECG App and PulseAI Algorithm	Inadequate reporting relating to Apple Watch metric. Study primarily focuses on PulseAI algorithm.
Dooley et al., (2017)	Estimating accuracy at exercise intensities: A comparative study of self-monitoring heart rate and physical activity wearable devices	Apple Watch not worn in accordance with manufacturer guidelines. Inappropriate criterion (Actigraph) for physical activity.
Dreisbach et al., (2023)	VALIDATION OF ACTIVITY MONITORS IN AND OUT-OF-THE LABORATORY ENVIRONMENT	Inadequate reporting (grey literature).
Ferrara et al., (2017)	Caloric Expenditure Using Indirect Calorimetry, Apple Watch Sport, and Fitbit Zip.	Inadequate reporting of statistical analysis.
Fiorina et al., (2021)	An artificial intelligence-based Smartwatch ECG analysis improves the detection of atrial arrhythmia	Inadequate reporting of statistical analysis and incorrect study objective. Study primarily assesses AI-based analysis.
Fokkema et al., (2017)	Reliability and Validity of Ten Consumer Activity Trackers Depend on Walking Speed.	Three wearable devices worn concurrently on the same wrist.
Ford et al., (2022)	Comparison of 2 Smart Watch Algorithms for Detection of Atrial Fibrillation and the Benefit of Clinician Interpretation	Manual duplicate removal.

Ford et al., (2022)	Comparison of 2 Smart Watch Algorithms for Detection of Atrial Fibrillation and the Benefit of Clinician Interpretation	Manual duplicate removal.
Garabelli et al., (2017)	First validation of an Apple® watch ECG event recorder	Inadequate reporting.
Ge et al., (2016)	Evaluating the accuracy of wearable heart rate monitors	Inadequate reporting of statistical analysis in comparison of Apple Watch with the criterion. Mean ranges and graphs with large scales presented.
Ghaly et al., (2022)	Accuracy of Pulse Oximetry Using Smartwatch Technology in COVID-19 Patients	Inadequate reporting of protocol and criterion (grey literature).
Glinskii et al., (2021)	Feasibility and validation testing of consumer wearables-based platform for remote monitoring of the six minute walk test in pulmonary arterial hypertension	Incorrect study objective. Primarily assesses the WalkTalkTrack app.
Gupta et al., (2023)	A comparison between the accuracy of apple watch 4 and Fitbit versa to measure heart rate	A conference abstract with insufficient reporting of methods, particularly of criterion measurement and statistical analysis.
Hajj-Boutros et al., (2023)	Wrist-worn devices for the measurement of heart rate and energy expenditure: A validation study for the Apple Watch 6, Polar Vantage V and Fitbit Sense	Multiple wearable devices worn concurrently during validation. Authors state that no manufacturer guidelines for device placement are provided by Apple, however, this is not the case.
Halsey G. (2025)	Smart Ring Outperforms Apple Watch for Atrial Fibrillation Detection.	Incorrect study objective. This is a summary of a separate research article.
Heinz et al., (2025)	Accuracy of energy expenditure estimation by the Apple Watch in EMS-supported exercise.	Inadequate reporting – particularly of the method for obtaining energy expenditure estimates with Apple Watch.
Held et al., (2022)	Agreement of the Apple Watch® and Fitbit Charge® for recording step count and heart rate when exercising in water	Step count in water – inappropriate metric.
Helmer et al., (2022)	Accuracy and Systematic Biases of Heart Rate Measurements by Consumer-Grade Fitness Trackers in Postoperative Patients: Prospective Clinical Trial	Several wrist-worn devices worn simultaneously on one wrist. Apple Watch not worn in accordance with manufacturer guidelines.

Hernando et al., (2018)	Validation of the Apple Watch for Heart Rate Variability Measurements during Relax and Mental Stress in Healthy Subjects	Inappropriate statistical analysis. 'Synchronization' between Apple Watch and Polar RR intervals was conducted – post-hoc gaps inserted. Inadequate detail in presentation of results.
Inocian et al., (2024)	Accuracy of the Apple Watch in detecting atrial fibrillation among patients undergoing 24-hour Holter monitoring	Inadequate reporting (grey literature).
Inui et al., (2020)	Use of a Smart Watch for Early Detection of Paroxysmal Atrial Fibrillation: Validation Study	Two wrist-worn devices on one wrist. Apple Watch not worn in accordance with manufacturer guidelines.
Isenegger et al., (2024)	DIAGNOSTIC ACCURACY OF SMARTWATCHES FOR AUTOMATED ATRIAL FIBRILLATION DETECTION OVER TIME: INSIGHTS FROM THE BASEL WEARABLE STUDY	Grey literature – conference abstract version of a paper that was subsequently published.
Jaworski et al., (2023)	Apple Watch Sleep and Physiological Tracking Compared to Clinically Validated Actigraphy, Ballistocardiography and Polysomnography.	One participant, two devices on one wrist.
Kennedy et al., (2023)	The effects of advanced filtering methods on visualisation of single-lead smartwatch ECG waveforms for clinical interpretation	Incorrect study objective. Examines ECG filtering techniques.
Khushhal et al., (2024)	Accuracy of Apple Watch to Measure Cardiovascular Indices in Patients with Chronic Diseases: A Cross Sectional Study.	The authors subsequently published a second study, with the same title, in a separate journal with a larger sample size.
Kobel et al., (2022)	Accuracy of the Apple Watch iECG in Children With and Without Congenital Heart Disease	Inappropriate intervention. Apple Watch ECG not designed for users under 22 years.
Kooner et al., (2024)	Commercially available activity monitors such as the fitbit charge and apple watch show poor validity in patients with gait aids after total knee arthroplasty.	Several devices worn simultaneously on one wrist. Inadequate description of methods.

Kwon et al., (2021)	Validation of the Apple Watch for Estimating Moderate-to-Vigorous Physical Activity and Activity Energy Expenditure in School-Aged Children	Inadequate reporting. Insufficient description of method for calculating MVPA with Apple Watch.
LaMunion et al., (2020)	Use of consumer monitors for estimating energy expenditure in youth	Inadequate reporting. Insufficient detail and clarity provided for methods and statistical analysis, in particular.
Lee et al., (2017)	"How 'bout Them Apples?" Validating Step Counts From The Apple Watch	Inadequate reporting in this conference abstract for inclusion.
Lee et al., (2022)	Accuracy of Wearable Devices for Estimating Energy Expenditure and Heart Rate During Golf	This paper was first machine-translated to English. Inadequate description of how the wearable devices were worn and tested. Five devices listed, not specified that they were worn at different times.
Leroux et al., (2023)	Feasibility and Diagnostic Value of Recording Smartwatch Electrocardiograms in Neonates and Children	Apple Watch ECG not designed for users under 22 years.
Le et al., (2022)	Validity of three smartwatches in estimating energy expenditure during outdoor walking and running	Multiple wrist-worn devices worn simultaneously.
Littell et al., (2022)	ASSESSMENT OF APPLEWATCH SERIES 6 PULSE OXIMETRY AND ECG ALGORITHM IN CHILDREN	Apple Watch ECG not designed for users under 22 years. Inadequate reporting of methods and statistical analysis (grey literature).
Littell et al., (2022)	Assessment of Apple Watch Series 6 pulse oximetry and electrocardiograms in a pediatric population.	Inappropriate intervention, as above. For pulse oximetry, Apple Watch was placed in areas other than the wrist. Measurements not obtained as per manufacturer guidelines.
Lynn et al., (2020)	Step-Counting Validity of Wrist-Worn Activity Monitors During Activities With Fixed Upper Extremities	Validating step count without arms swing. Measurements not obtained as per manufacturer guidelines.
Mak et al., (2020)	Reliability and repeatability of a smartphone-based 6-min walk test as a patient-centred outcome measure	Assesses step count from iPhone rather than Apple Watch.
Martín-Escudero et al., (2023)	Are Activity Wrist-Worn Devices Accurate for Determining Heart Rate during Intense Exercise?	Two devices worn on one wrist, as per photo included in study. No description of Apple Watch model.

Miller et al., (2022)	A Validation of Six Wearable Devices for Estimating Sleep, Heart Rate and Heart Rate Variability in Healthy Adults	Four wrist-worn devices worn simultaneously – Apple Watch not worn in accordance with manufacturer guidelines.
Moayedi et al., (2017)	Assessing the use of wrist-worn devices in patients with heart failure: feasibility study	Inadequate reporting of intervention and statistical analysis.
Mora-Gonzalez et al., (2022)	A catalog of validity indices for step counting wearable technologies during treadmill walking: the CADENCE-adults study	Inadequate description of methods and results. Unclear how many participants wore Apple Watch. Uninterpretable figures provided in results.
Muhonen et al., (2023)	Audiometric Validation of a Smart Watch Decibel Meter	Inappropriate validation protocol. Apple Watch placed 50 cm from a speaker in an audiometric booth, but not actually worn on the wrist.
Nasarre et al., (2022)	Using a smartwatch electrocardiogram to detect abnormalities associated with sudden cardiac arrest in young adults	Inappropriate metric. This conference abstract investigates detection of sudden cardiac arrest-associated abnormalities by clinician analysis of Apple Watch ECG recordings.
Nash et al., (2021)	But for the Blind Spot: Accuracy and diagnostic performance of the Apple Watch™ cardiac features in pediatric patients	Apple Watch ECG not designed for users under 22 years.
Paech et al., (2022)	Accuracy of the Apple Watch single-lead ECG recordings in pre-term neonates	Apple Watch ECG not designed for users under 22 years.
Paech et al., (2022)	Accuracy of the Apple Watch single-lead ECG recordings in pre-term neonates	Duplicate.
Park et al., (2021)	Free-Living Validation and Harmonization of 10 Wearable Step Count Monitors	Inappropriate criterion measure – no direct observation or manual counting. Comparison of Apple Watch to another wrist-worn wearable.
Park et al., (2021)	Free-living validation and harmonization of 10 wearable step-count monitors. 2021 ACSM Annual Meeting & World Congresses [Virtual], June 1 -5, 2021	Duplicate.

Park et al., (2021)	Free-living validation and harmonization of 10 wearable step-count monitors.	Duplicate.
Pasli et al., (2024)	Diagnostic accuracy of apple watch ECG outputs in identifying dysrhythmias: A comparison with 12-Lead ECG in emergency department.	Inadequate description of statistical analysis and inadequate detail in presentation of results.
Pätz et al., (2023)	Accuracy of the Apple Watch Oxygen Saturation Measurement in Adults and Children with Congenital Heart Disease	SpO_2 measurement with Apple Watch is not designed for use in children or neonates. Initially, authors did not tie the Apple Watch band, which led to many poor measurements. Inadequate detail, and separation of results, provided for the measurements taken when the study's methodology was altered midway through.
Perino et al., (2021)	Arrhythmias Other Than Atrial Fibrillation in Those With an Irregular Pulse Detected With a Smartwatch	Physician-adjudicated Apple Watch ECGs for the presence of atrial arrhythmia other than atrial fibrillation. This is not natively available to users.
Piccini et al., (2024)	Performance of apple watch for detecting atrial fibrillation among a known atrial fibrillation population	Inadequate description of criterion measure and statistical analysis (grey literature). Of 97 participants, only 20 included in final analysis.
Pipek et al., (2021)	Comparison of SpO2 and heart rate values on Apple Watch and conventional commercial oximeters devices in patients with lung disease	Finger oximeter is an inappropriate criterion method for heart rate. Authors provide no reference for the validity of either pulse oximeter for SpO_2 measurement. Our authorship team was unable to find any independent validation studies, or documentation showing regulatory clinical validation.
Ploux et al., (2022)	Beyond the wrist: Using a smartwatch electrocardiogram to detect electrocardiographic abnormalities	Inappropriate criterion measure. Interpretation of ECGs by clinician's rather than Apple Watch automated diagnosis.
Pope et al., (2019)	Validation of Four Smartwatches in Energy Expenditure and Heart Rate Assessment During Exergaming	Inappropriate criterion measure – Actigraph.

Pope et al., (2019)	Accuracy of commercially available smartwatches in assessing energy expenditure during rest and exercise	Inappropriate criterion measure – Actigraph.
Reece et al., (2021)	Accuracy Of Commercially Available Heart Rate Monitors Using Consumer Technology Association Standards...2021 ACSM Annual Meeting & World Congresses [Virtual], June 1 -5, 2021	Inadequate reporting (grey literature).
Reece et al., (2021)	Accuracy Of Commercially Available Heart Rate Monitors Using Consumer Technology Association Standards.	Inadequate reporting of results.
Rens et al., (2021)	Activity data from wearables as an indicator of functional capacity in patients with cardiovascular disease	Incorrect study objective. Assessment of app and 'frailty' was the primary objective.
Robbins et al., (2024)	Performance of Widely used Consumer Wearables for Sleep Tracking	Manual duplicate removal.
Roberts et al., (2020)	Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography	Multiple wearable devices worn concurrently on the same wrist.
Roomkham et al., (2018)	Can we use the apple watch to measure sleep reliably?	Inappropriate criterion measure – no polysomnography.
Sabatini et al., (2020)	QT Monitoring in Early Stage Asymptomatic/Paucisymptomatic Covid-19 Subjects on Hydroxychloroquine/Azithromycin Off-label Regimen - Iwatch ECG	Inadequate reporting of results and methods (grey literature).
Saghir et al., (2021)	Correlation of electrocardiographic rhythm interpretations in known atrial fibrillation or atrial flutter patients using the Apple Watch.	Conference abstract of a paper that was subsequently published.
Saghir et al., (2022)	Evaluation of Apple Watch electrocardiogram technology to detect atrial fibrillation in cardiac electrophysiology patients: comparing computer versus	Conference abstract of a paper that was subsequently published. Insufficient reporting of methods.

	physician interpretation in patients undergoing rhythm control strategy for atrial fibrillation.	
Samol et al., (2019)	Recording of Bipolar Multichannel ECGs by a Smartwatch: Modern ECG Diagnostic 100 Years after Einthoven	Assessing the ability of cardiologists of different experience levels to classify ECGs.
Sanchez et al., (2017)	Detecting atrial fibrillation using a smart watch - The mrhythm study	Inadequate reporting.
Sañudo et al., (2019)	Pilot Study Assessing the Influence of Skin Type on the Heart Rate Measurements Obtained by Photoplethysmography with the Apple Watch	Inadequate reporting of criterion measure and methods.
Savage et al., (2022)	Validation of Apple Watch heart rate monitoring for patients under general anaesthesia	Inadequate reporting of methods and results (grey literature).
Scholten et al., (2022)	Six-lead device superior to single-lead smartwatch ECG in atrial fibrillation detection	Inadequate description of statistical analysis and decision to exclude uninterpretable recordings from all analyses (20% of all tracings uninterpretable). Inadequate description of methods.
Scholten et al., (2021)	A comparison of over-the-counter available smartwatches and devices for electrocardiogram based detection of atrial fibrillation	Inappropriate methodology for validation. One measurement with ECG compared to four measurements from Apple Watch. All uninterpretable Apple Watch measurements then excluded from analyses. Inadequate description of methods.
Schramm et al., (2022)	Comprehensive cardiac interval analysis in the WEAR-TECH study cohort by comparing the Apple Watch Series 6 against a simultaneous 12-lead ECG	Inadequate reporting of statistical analysis and methodology (grey literature).
Schroeder et al., (2023)	The Apple Watch spO2 sensor and outliers in healthy users	Inadequate reporting of criterion measure and methodology.
Schyvens et al., (2025)	A performance validation of six commercial wrist-worn wearable sleep-tracking devices for sleep stage scoring compared to polysomnography	Multiple wearable devices (up to three) worn on the wrist simultaneously.

Sears et al., (2017)	Wrist-worn Physical Activity Trackers Tend to Underestimate Steps During Walking.	Inadequate reporting of methods and results.
Shakhtour et al., (2024)	Evaluation of Noise Exposure Levels in Pediatric ENT Operating Rooms.	Inappropriate criterion measure – iPhone application used for sound exposure.
Shcherbina et al., (2017)	Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort	Up to four devices worn simultaneously on the wrist. Apple Watch not worn in accordance with manufacturer guidelines.
Singh et al., (2019)	Expecting the unexpected beyond borders: Apple watch as a global screening tool for dysrhythmia in the developing world	Inadequate description of methodology.
Spaccarotella et al., (2020)	Validation of Remote Measurement of the Qtc Intervals Using an Apple Watch	Inadequate description of methodology.
Spaccarotella et al., (2020)	Multichannel electrocardiograms obtained by a smartwatch for the diagnosis of the acute coronary syndromes: the smart AMI trial	Manual duplicate removal.
Spaccarotella et al., (2020)	Multichannel electrocardiograms obtained by a smartwatch for the diagnosis of acute coronary syndromes: The smart ami trial	Manual duplicate removal.
Sprenger et al., (2022)	Feasibility and Reliability of Smartwatch to Obtain Precordial Lead Electrocardiogram Recordings	ECG not taken with Apple Watch worn on the wrist.
Stove et al., (2023)	Assessment of Noninvasive Oxygen Saturation in Patients With COPD During Pulmonary Rehabilitation: Smartwatch versus Pulse Oximeter	Inappropriate criterion measure. Pulse oximeter not medical-grade or validated.
Striepe et al., (2022)	Use of the Apple Watch iECG in adult congenital heart disease patients.	Manual duplicate removal.
Teich et al., (2023)	Development of an AI based automated analysis of pediatric Apple Watch iECGs	ECG not designed for use in individuals under 22 years of age.

Thomson et al., (2018)	Validity of Heart Rate Measurements for the Apple Watch and Fitbit Charge HR 2	Inadequate description of results.
Toner et al., (2019)	The Accuracy of Smartwatches Compared to Holter Monitors for Heart Rate Monitoring in Atrial Fibrillation: A Pilot Study	Inadequate description of methodology (grey literature).
Tran et al., (2024)	Peak oxygen consumption by smartwatches compared with cardiopulmonary exercise test in complex congenital heart disease.	Inadequate description of cardiopulmonary exercise testing procedure, and results.
Tran et al., (2023)	Peak oxygen consumption by smartwatches compared with cardiopulmonary exercise test in complex congenital heart disease	Manual duplicate removal.
Turki et al., (2021)	Estimation of Heart Rate Variability Measures Using Apple Watch and Evaluating Their Accuracy: Estimation of Heart Rate Variability Measures Using Apple Watch	State the Apple Watch heart rate recordings were interpolated and resampled, however there is inadequate description of this process and its rationale. Inadequate description of statistical analysis.
Van der Zande et al., (2023)	Using a smartwatch to record precordial electrocardiograms: A validation study	The primary objective of this study was to assess the validity of Apple Watch to measure precordial ECGs, at various locations around the body, which Apple Watch is not designed to do. At these various locations, Apple Watch was held in place against the skin. Based on the description provided, Apple Watch was also held against the skin at the wrist for Lead I, rather than worn with the strap closed.
Van der Zande et al., (2023)	Using a smartwatch to record precordial electrocardiograms: A validation study	Manual duplicate removal.
Veerabhadrappa et al., (2018)	Fitness trackers may benefit cardiovascular health, but are they accurate?	Inadequate description of methodology.
Viciana et al., (2022)	Are Wrist-Worn Activity Trackers and Mobile Applications Valid for	Apple Watch not worn in accordance with manufacturer guidelines.

	Assessing Physical Activity in High School Students? Wearfit Study	Multiple wrist-worn devices on one wrist.
Wasserlauf et al., (2018)	Accurate detection and quantification of a-trial fibrillation using a smart-watch with ECG watch band	Assessment of the KardiaBand paired with early generation of Apple Watch.
Wasserlauf et al., (2019)	Smartwatch Performance for the Detection and Quantification of Atrial Fibrillation	Assessment of the KardiaBand paired with early generation of Apple Watch.
Weaver et al., (2023)	Evaluation of a device-agnostic approach to predict sleep from raw accelerometry data collected by Apple Watch Series 7, Garmin Vivoactive 4, and ActiGraph GT9X Link in children with sleep disruptions	Inappropriate intervention.
Weaver et al., (2023)	A Device Agnostic Approach to Predict Children's Activity from Consumer Wearable Accelerometer Data: A Proof-of-Concept Study	Inappropriate intervention – metric not natively available on Apple Watch.
Weidlich et al., (2023)	Reducing the Burden of Inconclusive Smart-Device Single-lead ECG tracings via a Novel Artificial Intelligence Algorithm	Inappropriate intervention. Medical professionals diagnosing arrhythmia based on Apple Watch ECG.
Weidlich et al., (2023)	Impact of clinical expertise and choice of wearable device on accuracy to detect atrial fibrillation via single-lead ECGs	Incorrect study objective.
Wexler et al., (2021)	Estimating basal metabolic rate in people with diabetes	Incorrect study objective.
White et al., (2023)	Comparison of raw accelerometry data from ActiGraph, Apple Watch, Garmin, and Fitbit using a mechanical shaker table	Raw accelerometry data rather than health-related metric.
Xie et al., (2018)	Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: Comparative study	Apple Watch not worn in accordance with manufacturer guidelines. Multiple wrist-worn devices worn simultaneously on one wrist.

Yalin et al., (2023)	Diagnostic accuracy of Apple Watch Series 6 recorded single-lead ECGs for identifying supraventricular tachyarrhythmias: a comparative analysis with invasive electrophysiological study	Incorrect study objective. Assessing the ability of clinician's to assess Apple Watch ECG recordings.
Yee-ming Li et al., (2023)	Detection of QT interval prolongation using Apple Watch electrocardiogram in children and adolescents with congenital long QT syndrome	Apple Watch ECG recording designed for use in adults aged 22 years and older. QTc interval was interpreted by blinded investigators.
Zahedivash et al., (2023)	Utility of the Apple Watch® for identifying arrhythmias in children	Apple Watch ECG not designed for individuals under 22 years of age.
Zhang et al., (2017)	Validating iWatch in Measuring Energy Expenditure during Different Levels of Physical Activity	Conference abstract of a paper that was subsequently published and evaluated separately as part of our study.

References

1 Støve, M. P. & Hansen, E. C. K. Accuracy of the Apple Watch Series 6 and the Whoop Band 3.0 for assessing heart rate during resistance exercises. *Journal of sports sciences* 40, 2639-2644 (2022).

2 Gillinov, S. *et al.* Variable accuracy of wearable heart rate monitors during aerobic exercise. *Medicine & Science in Sports & Exercise* 49, 1697-1703 (2017).

3 Reece, J. D., Bunn, J. A., Choi, M. & Navalta, J. W. Assessing heart rate using consumer technology association standards. *Technologies* 9, 46 (2021).

4 O'Grady, B., Lambe, R., Baldwin, M., Acheson, T. & Doherty, C. The Validity of Apple Watch Series 9 and Ultra 2 for Serial Measurements of Heart Rate Variability and Resting Heart Rate. *Sensors (Basel)* 24 (2024). https://doi.org/10.3390/s24196220

5 Mannhart, D. *et al.* Clinical validation of an artificial intelligence algorithm offering cross-platform detection of atrial fibrillation using smart device electrocardiograms. *Archives of Cardiovascular Diseases* 116, 249-257 (2023).

6 Mannhart, D. *et al.* Clinical validation of 5 direct-to-consumer wearable smart devices to detect atrial fibrillation: BASEL wearable study. *Clinical Electrophysiology* 9, 232-242 (2023).

7 Racine, H.-P. *et al.* Role of coexisting ECG anomalies in the accuracy of smartwatch ECG detection of atrial fibrillation. *Canadian Journal of Cardiology* 38, 1709-1712 (2022).

8 Scholten, J. *et al.* A comparison of over-the-counter available smartwatches and devices for electrocardiogram based detection of atrial fibrillation. *European Heart Journal-Digital Health* 2, ztab104. 3047 (2021).

9 Velraeds, A. *et al.* Improving automatic smartwatch electrocardiogram diagnosis of atrial fibrillation by identifying regularity within irregularity. *Sensors* 23, 9283 (2023).

10 Seshadri, D. R. *et al.* Accuracy of Apple Watch for detection of atrial fibrillation. *Circulation* 141, 702-703 (2020).

11 Perez, M. V. *et al.* Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine* 381, 1909-1917 (2019).

12 Jiang, Y. *et al.* Performance of Wearable Pulse Oximetry During Controlled Hypoxia Induction. (2024). https://doi.org/10.1101/2024.07.16.24310506

13 Helmer, P. *et al.* Reliability of continuous vital sign monitoring in post-operative patients employing consumer-grade fitness trackers: A randomised pilot trial. *Digit Health* 10, 20552076241254026 (2024). https://doi.org/10.1177/20552076241254026

14 Rafl, J., Bachman, T. E., Rafl-Huttova, V., Walzel, S. & Rozanek, M. Commercial smartwatch with pulse oximeter detects short-time hypoxemia as well as standard medical-grade device: Validation study. *Digital health* 8, 20552076221132127 (2022).

15 Rajakariar, K. *et al.* Accuracy of Smartwatch Pulse Oximetry Measurements in Hospitalized Patients With Coronavirus Disease 2019. *Mayo Clinic Proceedings: Digital Health* 2, 152-158 (2024). https://doi.org/10.1016/j.mcpdig.2024.02.001

16 Walzel, S. *et al.* Evaluation of leading smartwatches for the detection of hypoxemia: comparison to reference oximeter. *Sensors* 23, 9164 (2023).

17 O'Driscoll, B. R., Howard, L., Earis, J. & Mak, V. BTS guideline for oxygen use in adults in healthcare and emergency settings. *Thorax* 72, ii1-ii90 (2017).

18 Fischer, T., Schraivogel, S., Caversaccio, M. & Wimmer, W. Are smartwatches a suitable tool to monitor noise exposure for public health awareness and otoprotection? *Frontiers in neurology* 13, 856219 (2022).

19 Dreisbach, S., Rhudy, M., Moran, M., Henriquez, B. & Veerabhadrappa, P. Accuracy of Apple Watch and Actigraphs during overground and treadmill walking. *HUMAN MOVEMENT* 26, 83-90 (2025). https://doi.org/10.5114/hm/202462